Claims:

1. A method of extracting new word automatically, said method comprising the steps of:

segmenting a cleaned corpus to form a segmented corpus;

splitting the segmented corpus to form sub strings, and counting the occurrences of each sub strings appearing in the corpus; and

filtering out false candidates to output new words.

2. The method of extracting new word automatically according to Claim 1, wherein the step of segmenting comprises using punctuations, Arabic digits and alphabetic strings, or new words patterns to split the cleaned corpus.

3. The method of extracting new word automatically according to Claim 1, wherein the step of segmenting comprises using common vocabulary to segment the cleaned corpus.

4. The method of extracting new word automatically according to Claim 1, wherein the step of splitting and counting is implemented using a GAST.

5. The method of extracting new word automatically according to Claim 4, wherein a GAST is implemented by limiting length of sub strings.

6. The method of extracting new word automatically according to Claim 1, wherein the step of filtering out false candidates comprises:

filtering out functional words;

filtering out those sub strings which almost always appear along with a longer sub strings; and

filtering out those sub strings for which the occurrence is less than a predetermined threshold.

7. The method of extracting new word automatically according to Claim 1, wherein the step of segmenting the cleaned corpus comprises using pre-recognized functional words as segment boundary patterns.

8. The method of extracting new word automatically according to Claim 3, wherein the step of segmenting cleaned corpus comprises using pre-recognized functional words as segment boundary patterns.

9. The method of extracting new word automatically according to Claim 3, wherein the step of filtering out false words comprises:

filtering out functional words;

filtering out those sub strings which almost always appear along with a longer sub strings; and

filtering out those sub strings for which the occurrence is less than a predetermined threshold.

10. An automatic new word extraction system, comprising:

a segmentor which segments a cleaned corpus to form a segmented corpus;

a splitter which splits the segmented corpus to form sub strings, and which counts the number of the sub strings appearing in the corpus; and

a filter which filters out false candidates to output new words.

11. The automatic word extraction system according to Claim 10, wherein the segmentor uses punctuations, Arabic digits and alphabetic strings, or new word pattern to segment the cleaned corpus.

12. The automatic word extraction system according to Claim 10, wherein the segmentor uses common vocabulary to segment the cleaned corpus.

13. The automatic word extraction system according to Claim 10, wherein the splitter builds a GAST.

14. The automatic word extraction system according to Claim 13, wherein the GAST limits the length of sub strings.

15. The automatic word extraction system according to Claim 10, wherein the filter filters out functional words; those sub strings which almost always appear along with longer sub strings; and those sub strings for which the occurrence is less than a predetermined threshold.

16. The automatic word extraction system according to Claim 10, wherein the segmentor uses pre-recognized functional words as segment boundary patterns.

17. The automatic word extraction system according to Claim 12, wherein the segmentor uses pre-recognized functional words as segment boundary patterns.

18. The automatic word extraction system according to Claim 12, wherein the filter filters out functional words; those sub strings which almost always appear along with a longer sub strings; and those sub strings for which the occurrence is less than a predetermined threshold.

19. A program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for extracting new word automatically, said method comprising the steps of:

segmenting a cleaned corpus to form a segmented corpus;

splitting the segmented corpus to form sub strings, and counting the occurrences of each sub strings appearing in the corpus; and

filtering out false candidates to output new words.